International Journal of Epidemiology, 2020, 1–10
doi: 10.1093/ije/dyaa192
Original Article

Original Article

Factors associated with participation over time in the Avon Longitudinal Study of Parents and Children: a study using linked education and primary care data

Rosie P Cornish,[1,2]* John Macleod,[1] Andy Boyd[1] and Kate Tilling[1,2]

[1]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK and [2]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK

*Corresponding author. Population Health Sciences, University of Bristol, Oakfield House, Oakfield Grove, Bristol BS8 2BN, UK. E-mail: rosie.cornish@bristol.ac.uk

Editorial decision 10 August 2020; Accepted 10 September 2020

Abstract

Background: In observational research, choosing an optimal analysis strategy when variables are incomplete requires an understanding of the factors associated with ongoing participation and non-response, but this cannot be fully examined with incomplete data. Linkage to external datasets provides additional information on those with incomplete data, allowing examination of factors related to missingness.

Methods: We examined the association between baseline sociodemographic factors and ongoing participation in the Avon Longitudinal Study of Parents and Children. We investigated whether child and adolescent outcomes measured in linked education and primary care data were associated with participation, after accounting for baseline factors. To demonstrate the potential for bias, we examined whether the association between maternal smoking and these outcomes differed in the subsample who completed the 19-year questionnaire.

Results: Lower levels of school attainment, lower general practitioner (GP) consultation and prescription rates, higher body mass index (BMI), special educational needs (SEN) status, not having an asthma diagnosis, depression and being a smoker were associated with lower participation after adjustment for baseline factors. For example, the adjusted odds ratio (OR) for participation comparing ever smokers (by 18 years) with non-smokers was: 0.65, 95% CI (0.56, 0.75). The associations with maternal smoking differed between the subsample of participants at 19 years and the entire sample, although differences were small and confidence intervals overlapped. For example: for SEN status, OR = 1.19 (1.06, 1.33) (all participants); OR = 1.03 (0.79, 1.45) (subsample).

Conclusions: A range of health-related and educational factors are associated with ongoing participation in ALSPAC; this is likely to be the case in other cohort studies.

Researchers need to be aware of this when planning their analysis. Cohort studies can use linkage to routine data to explore predictors of ongoing participation and conduct sensitivity analyses to assess potential bias.

Key words: ALSPAC, non-response, participation, data linkage, selection bias, missing data

Key Messages

- Educational and health-related characteristics are strongly associated with ongoing participation in the Avon Longitudinal Study of Parents and Children, after adjustment for detailed sociodemographic factors.
- This could bias analyses using the dataset, with bias dependent on the variables used in the analysis and their impact on participation.
- Examination of factors associated with participation is important when analysing partially observed data in order to assess potential bias, but this cannot generally be done using study data alone.
- Researchers can use linkage to external sources of data to identify such factors, to make informed decisions about the likely impact of selective participation and to inform their analyses.
- Similar patterns of non-response are likely to be observed in other cohort studies, although the specific factors associated with participation may differ. Knowledge of these factors is important when comparing and summarising evidence across studies.

Introduction

Non-response and dropout from longitudinal studies (we will refer to not dropping out/response as participation) result in missing study information. This will bias estimates of exposure-outcome associations if the association between the exposure and the probability of participation is differential with respect to the outcome, or vice versa.[1] The circumstances under which this occurs depend on the analysis model being used.[2] To know whether a given analysis model is likely to be biased by dropout, we need to be able to assess whether the exposure and outcome are related to participation—this is impossible to do using only the observed study data. However, complete data on some post-baseline variables may be obtained via linkage to external datasets. As with any observational study, associations between a factor and participation may be causal (smoking may cause dropout) or confounded (lower education levels may cause both dropout and smoking). One way to address this is by using genetic information.[3,4] A recent study using Avon Longitudinal Study of Parents and Children (ALSPAC) data[3] found that, in both the mothers and the index children, polygenic scores for years of education and agreeableness were associated with greater participation; conversely, polygenic scores for smoking initiation, schizophrenia, attention-deficit hyperactivity disorder (ADHD) and depression were associated with lower levels of participation. However, using genetic information has several disadvantages. First, the genetic contribution to phenotypes is often low, limiting power; second, not all factors of interest can be genetically instrumented; and finally, genetic information may not be available for everyone.

Previous work in ALSPAC has shown that linked educational data allow examination of the missing data mechanism for intelligence quotient (IQ) and provide useful auxiliary variables for multiple imputation, leading to reduced bias in estimates of the association between breastfeeding and IQ.[5] Other studies have also used variables from external sources of data in inverse probability weighting or multiple imputation models, in order to reduce bias due to missing data in specific covariate or outcome data.[6–8]

Here we aimed to complement and extend previous studies by examining associations of participation with a range of observed phenotypes as recorded in linked education or primary care data, including factors that cannot be (or have not been) genetically instrumented [school absence, special educational needs classification, general practitioner (GP) consultation rates, counts of prescribed drugs]. We examine factors associated with continuing participation in ALSPAC, but the methods are applicable to any cohort study or long-term follow-up of a clinical trial.

Methods

Sample/subjects

ALSPAC is a prospective study—described in detail previously[9,10]—which recruited pregnant women living in and around Bristol, south-west England, with due dates between April 1991 and December 1992. Initially 14 541 pregnant women enrolled, resulting in 14 062 live births and 13 988 infants alive at 1 year. Detailed data were collected during pregnancy, and participants have been followed up since birth through questionnaires, clinics and linkage to routine datasets. (Note that ALSPAC has a searchable data dictionary and variable search tool: http://www.bristol.ac.uk/alspac/researchers/our-data/). We used data from singletons and twins who were alive at 1 year and had not subsequently withdrawn ($n = 13\,972$).

ALSPAC participation

ALSPAC suffers from attrition as well as sporadic non-response, with individuals participating at some time points but not others. We defined participation (at multiple time points) as returning a completed questionnaire or attending a study assessment clinic. (Note that a partially completed questionnaire in which some questions were left unanswered would be considered 'completed' in this context). Participants in this study were mothers or carers (henceforth termed mothers) who completed questionnaires about themselves or their child, and the (index) children who (from age 65 months) also completed questionnaires. Clinics before age 7 years were aimed at a subset of enrolled children; we only considered questionnaires and clinics (up to age 20) where all participants were eligible to complete/attend. We have considered mother-completed questionnaires (about themselves or their child) separately from clinics and child-completed questionnaires; the latter two constitute child participation, the former mother participation. A complete list of questionnaires and clinics that we included is given in Supplementary Table S1, available as Supplementary data at *IJE* online. This gave 42 binary participation measures for mothers and 33 binary participation measures for the children.

Baseline sociodemographic variables

Baseline sociodemographic and other variables potentially associated with participation were included. The majority of these were factors measured in pregnancy, since this was when response rates were highest. The variables were: maternal ethnicity, age, parity, marital status, age at first pregnancy, educational level, smoking and depression score (Edinburgh Postnatal Depression Scale at 18 weeks of gestation); housing tenure; whether the mother or her partner had use of a car; whether their house had double glazing; whether they had a phone in their home; number of rooms in the house; crowding index, defined as number of people per room (excluding bathrooms and toilets); family occupational social class, defined as the higher of maternal and paternal social class and categorized as I-IIIN (professional, managerial and non-manual skilled occupations) and IIIM-IV (manual skilled, semi-skilled and unskilled occupations); duration of breastfeeding, derived from responses to questionnaires administered at 4 weeks, 6 months and 15 months; and child sex. Paternal factors were not included—because response rates for these were lower and because, in a preliminary analysis, none of the paternal factors considered (education, smoking and depression score) were associated with participation, after taking account of maternal and other factors listed above.

Education variables from the National Pupil Database (NPD)

The NPD is a database containing attainment and other data for children attending schools in England (https://www.gov.uk/government/collections/national-pupil-database). Linkage between ALSPAC and the NPD has been described previously.[5] Here we used three variables from Year 11 (age 15–16 years): capped point score (a measure of attainment), described previously,[5] percent attendance and special educational needs (SEN) status, classified as none, school action support or SEN Statement. [SEN Statements—now replaced by Education, Health and Care Plans (EHC Plans)—are a description of a child's educational needs and any additional support they should receive in school.]

Linkage to primary care data

When the index children reached legal adulthood (age 18), ALSPAC conducted a postal fair-processing campaign to re-enrol them into the study in their own right and to seek (opt-out) permission for linkage to health and administrative records. Linkage to primary care records was carried out following this campaign and is described in the Supplementary Material, available as Supplementary data at *IJE* online (Section 2).

Variables derived from linked primary care data

- Adolescent body mass index (BMI): the mean of all recorded measurements after age 10;

- consultation rate age 15–19 years: the total number of consultations during this period, divided by five;
- mean (prescribed) drug count age 15–19 years: total drug count in this period divided by five;
- asthma diagnosis: Read code for a diagnosis[11] before 8 years;
- depression: Read code for a diagnosis, symptoms or treatment before the age of 18[12];
- smoking: based on Read codes used in two recent studies,[13,14] we classified individuals as having a record for ever or current smoking (or not) before age 18 (Supplementary Material, Section 3, available as Supplementary data at *IJE* online).

Numbers with linked data

Of the 13 972 individuals included in this analysis, 12 395 (89%) were sent fair-processing materials. Of these, 360 (3%) dissented to linkage to education records and 423 (3%) to health records. Of the remaining 12 035, 11 414 were linked to the NPD and had data on at least one of the variables used in this analysis. Of the 11 972 who did not dissent to health data linkage, ALSPAC had no National Health Service (NHS) ID for 17, leaving 11 955 where linkage to primary care records was attempted. Primary care records (not necessarily for the entire time period) were extracted for 11 087 (93% of individuals where linkage was possible; 79% of the original 13 972).

Statistical analysis

For each questionnaire and study clinic, a binary variable was created to indicate whether each individual participated (returned that questionnaire or attended the clinic). Two random effects logistic regression models (one for mother and one for child participation) were used to model participation over time, using cubic splines[15] with five knots placed using Stata's default method.[16] We used the (fixed) age at which the questionnaire/clinic invitation was sent, rather than the actual age at completion/attendance. For the mother-completed questionnaires, we used time in the study, where time in the study = 0 denotes the beginning of pregnancy.

Multiple imputation using chained equations[17] was used to impute missing data. Two models were used. The first imputation model (which included all 13 972 individuals) was used to impute baseline covariates and linked education variables; the model included the baseline covariates, the three linked education variables (attainment score, percent absence and SEN status) and all binary participation variables (both mother and child). IQ measured when the children were aged 8 using the Wechsler Intelligence Scale for Children (WISC-III)[18] was included as an auxiliary variable. A second imputation model was used to examine the association between measures based on primary care data and participation. This model only included individuals with primary care data beyond the age of four ($n = 10\ 811$) and used baseline covariates, all participation variables, the seven GP measures and, as auxiliary variables, consultation rates and drug counts at additional ages (0–4 years, 5–9 years, 10–14 years and 20+ years). Further details regarding missing data and the imputation models are given in the Supplementary Material (Sections 4 and 5; Supplementary Tables S2 and S3; available as Supplementary data at *IJE* online). As a sensitivity analysis, we carried out a complete case analysis.

To investigate the impact of selective participation on exposure-outcome estimates, we used linear and logistic regression (as appropriate) to examine the association between maternal smoking (ever vs never, measured in early pregnancy) and the following outcomes: attainment score, percent school absence, SEN status (dichotomized: school action support/statement of SEN vs none), asthma, BMI, depression and smoking. The numerical variables—attainment, school absence and BMI—were converted to z-scores [standard deviation (SD) units] for this analysis. These associations were examined (in the imputed data) among (i) all individuals and (ii) among individuals who completed the ALSPAC questionnaire administered at 19 years (the latest child-completed questionnaire included in our analysis). As a sensitivity analysis, this was repeated among the complete cases (i.e. all those with data on maternal smoking, the outcome of interest, plus baseline covariates) and then among the subset of these who completed the age 19 questionnaire. Analyses were carried out in Stata 15.[19]

Ethics approval

Ethical approval was obtained from the ALSPAC Ethics and Law Committee and local research ethics committees (http://www.bristol.ac.uk/alspac/researchers/research-ethics/).

Results

Participation rates by the mothers were high in pregnancy and gradually declined over time, particularly when the children reached mid-late adolescence (Figure 1). Overall rates of child participation remained stable during childhood (children started completing questionnaires at around 5 years of age) but declined during adolescence, although previous assessments have shown that sporadic

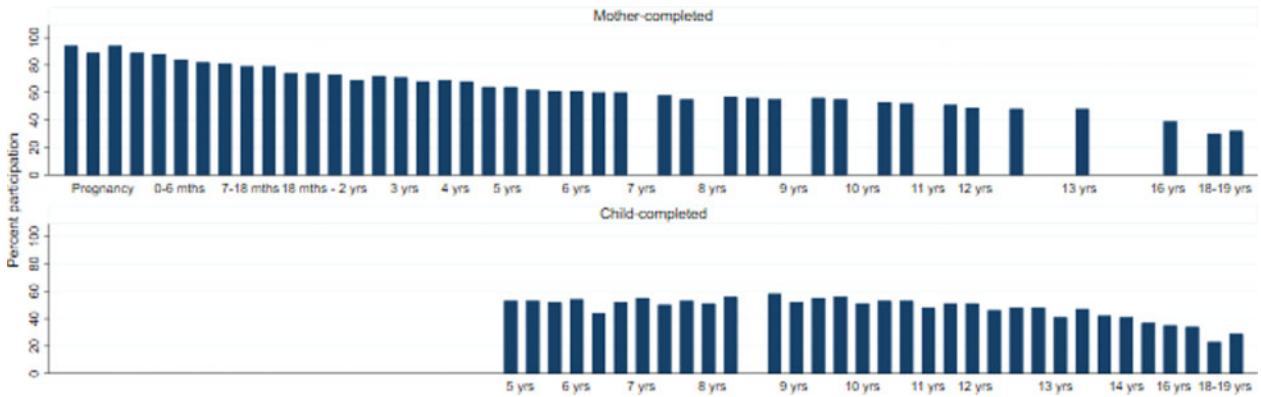

Figure 1 Participation rates (%) in the Avon Longitudinal Study of Parents and Children (ALSPAC): mother and child completed. Enrolment in ALSPAC is defined as the participant having actively participated at least once. However, no single assessment has complete coverage as there was no single baseline assessment administered to all participants, given that mothers were recruited at different stages of pregnancy or following delivery, and may therefore have missed an earlier assessment

Table 1 Odds ratios for participation for all baseline covariates (50 imputed datasets; $n = 13\,972$)

Covariate	Level	Child participation OR (95% CI)	*P*-value	Mother participation OR (95% CI)	*P*-value
Child sex	Female vs male	1.88 (1.71, 2.06)	<0.001	1.10 (0.99, 1.22)	0.07
Mother's education	O level/lower	1.00		1.00	
	A level	1.47 (1.28, 1.68)		1.58 (1.35, 1.85)	
	Degree/higher	1.66 (1.38, 2.00)	<0.001	1.74 (1.40, 2.18)	<0.001
Parity	0	1.00		1.00	
	1	0.81 (0.70, 0.93)		0.84 (0.72, 0.99)	
	2+	0.59 (0.48, 0.73)	<0.001	0.60 (0.48, 0.76)	<0.001
Mother's age (at birth of index child)	Per 1-year increase	1.07 (1.06, 1.09)	<0.001	1.07 (1.05, 1.08)	<0.001
Mother's ethnicity	Non-White vs White	0.26 (0.19, 0.36)	<0.001	0.14 (0.10, 0.19)	<0.001
Family social class	Manual vs non-manual	0.73 (0.63, 0.85)	<0.001	0.63 (0.52, 0.75)	<0.001
Age at first pregnancy	<20	1.00		1.00	
	20-24	1.36 (1.17, 1.59)		1.32 (1.11, 1.56)	
	25+	1.63 (1.37, 1.95)	<0.001	1.85 (1.51, 2.25)	<0.001
Maternal smoking	Yes vs no (in pregnancy)	0.80 (0.69, 0.93)	0.003	0.80 (0.67, 0.94)	0.008
	Yes vs no (ever)	0.78 (0.68, 0.88)	<0.001	0.81 (0.71, 0.93)	0.003
Duration of breastfeeding	Never/<1 month	1.00		1.00	
	1 to <3 months	2.99 (2.56, 3.51)		3.74 (3.12, 4.48)	
	3 to <6 months	2.69 (2.26, 3.20)		2.92 (2.37, 3.61)	
	6 months+	3.97 (3.45, 4.56)	<0.001	5.27 (4.46, 6.23)	<0.001
Married	Yes vs no	1.33 (1.16, 1.53)	<0.001	1.39 (1.19, 1.62)	<0.001
Housing tenure	Owned/mortgaged	1.00		1.00	
	Private rented	0.50 (0.40, 0.63)		0.49 (0.39, 0.63)	
	Council/HA[a]/other	0.69 (0.59, 0.82)	<0.001	0.66 (0.55, 0.79)	<0.001
Number of rooms	Per 1-room increase	1.02 (0.97, 1.07)	0.5	1.00 (0.94, 1.06)	1.0
Phone in home	Yes vs no/incoming only	0.66 (0.54, 0.79)	<0.001	0.68 (0.55, 0.84)	<0.001
Car use	No vs yes	0.53 (0.44, 0.65)	<0.001	0.61 (0.49, 0.76)	<0.001
Double glazing	None vs full/partial	0.88 (0.79, 0.98)	0.02	0.86 (0.76, 0.96)	0.009
Financial difficulties	Per 1-unit increase	0.98 (0.96, 0.99)	0.008	0.97 (0.95, 0.99)	0.004
Crowding index	≤0.5	1.00		1.00	
	>0.5–0.75	0.93 (0.80, 1.08)		0.85 (0.72, 1.01)	
	>0.75–1	0.85 (0.70, 1.03)		0.79 (0.62, 0.99)	
	>1	0.75 (0.57, 1.00)	0.2	0.59 (0.43, 0.81)	0.06
Depression score	Per 1-unit increase	0.98 (0.97, 0.99)	0.004	0.98 (0.97, 0.99)	0.001

[a]HA, housing association.

Table 2 Odds ratios for participation for education variables (50 imputed datasets; $n = 13\,972$)

Covariate	Level	Child participation OR (95% CI)[a]	*P*-value	Mother participation OR (95% CI)[a]	*P*-value
Attainment score	For 10-point increase	1.06 (1.05, 1.07)	<0.001	1.05 (1.04, 1.06)	<0.001
SEN[b] status	None	1.00		1.00	
	School action	0.76 (0.64, 0.91)		0.88 (0.73, 1.06)	
	Statement	0.61 (0.44, 0.85)	0.002	0.91 (0.65, 1.27)	0.5
School absence	For 1-point increase in square root of % absence	0.89 (0.86, 0.92)	<0.001	0.89 (0.85, 0.92)	<0.001

[a]Mutually adjusted and adjusted for baseline factors.
[b]Special educational needs.

Table 3 Odds ratios for participation for GP[a]-derived child measures (50 imputed datasets: $n = 10\,811$)

Covariate		Child participation OR (95% CI)[a]	*P*-value	Mother participation OR (95% CI)[b]	*P*-value
Asthma diagnosis before age 8	Yes vs no	8.71 (6.40, 11.83)	<0.001	8.99 (6.34, 12.74)	<0.001
Smoking record before age 18	Yes vs no	0.65 (0.56, 0.75)	<0.001	0.70 (0.60, 0.82)	<0.001
Depression before age 18	Yes vs no	0.79 (0.64, 0.97)	0.02	0.87 (0.70, 1.08)	0.2
BMI[a]	Per 1kg/m^2	0.99 (0.98, 1.00)	0.09	0.99 (0.98, 1.00)	0.2
Consultation rate age 15-19	≤1 per year	1.00		1.00	
	>1–4 per year	1.33 (1.15, 1.54)		1.04 (0.89, 1.22)	
	>4 per year	1.60 (1.36, 1.89)	<0.001	1.12 (0.94, 1.33)	0.4
Drug count age 15-19	≤1 per year	1.00		1.00	
	>1–4 per year	1.27 (1.12, 1.43)		1.04 (0.92, 1.19)	
	>4 per year	1.40 (1.18, 1.65)	<0.001	1.07 (0.89, 1.29)	0.5

[a]GP: general practitioner; BMI: body mass index.
[b]Adjusted for baseline factors.

participation means that there is a larger pool of participating individuals than the response to any single assessment suggests.[9]

Baseline covariate data were missing for between 0% (sex and mother's age at birth) and 18% (family occupational social class) of the 13 972 individuals (Supplementary Table S2) and 9049 (65%) had complete covariate information. Individuals with complete baseline covariates had higher levels of participation than the overall sample (mean number of mother-completed questionnaires was 31, compared with 27 in the overall sample; mean number of child-completed questionnaires/clinics attended was 19, compared with 16 in the overall sample).

Predictors of participation: baseline variables

Table 1gives mutually adjusted odds ratios (ORs) for child and mother participation [obtained using multiple imputation (MI))] for all baseline covariates. ORs for child and mother participation were very similar, with the exception of sex (of child), which was a strong predictor of child participation but not of mother participation. The ORs from the complete case analysis were generally similar to those obtained using multiple imputation (MI) (Supplementary Table S4, available as Supplementary data at *IJE* online), although ORs for breastfeeding and maternal ethnicity were less extreme in the complete case analysis.

Predictors of participation: child education variables

After adjusting for baseline factors, all three education variables were associated with child participation (Table 2). Those with higher attainment scores, lower absence rates and no recorded SEN were more likely to participate. Attainment and absence (but not SEN status) were also associated with mother participation. Results were similar in

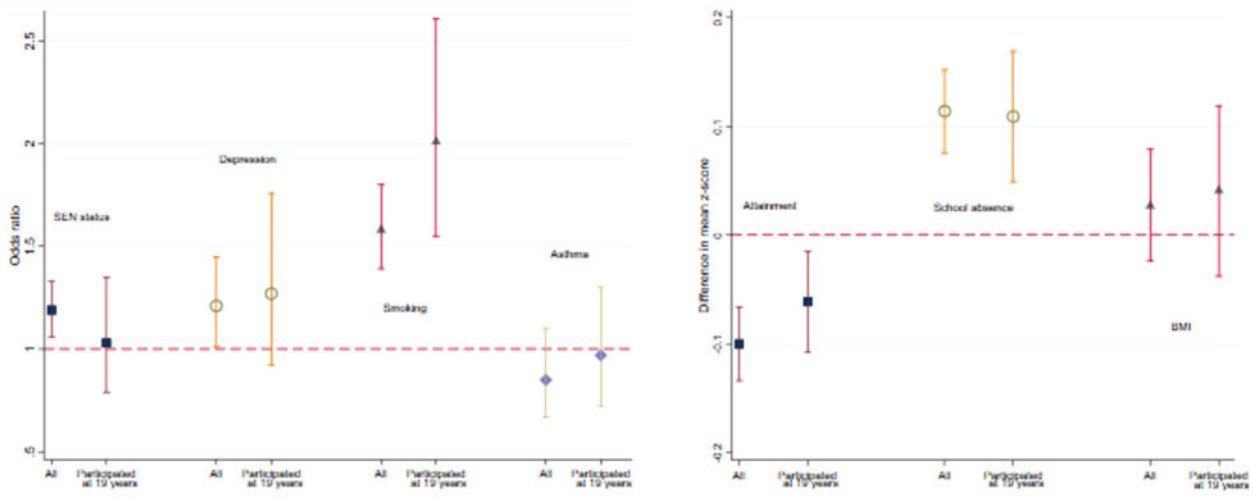

Figure 2 Association between maternal smoking (yes vs no) and various adolescent outcomes

the complete case analysis (Supplementary Table S5, available as Supplementary data at *IJE* online).

Predictors of participation: child measures derived from primary care data

After adjusting for baseline factors, all primary care measures were related to child participation, although the association with BMI was weak (Table 3). GP-recorded (child) asthma diagnosis by age 8 was strongly associated with increased participation by the mother and (GP-recorded child) smoking before age 18 was strongly associated with decreased participation by the mother. The ORs for the association between the other GP-derived variables and mother participation were in the same direction (but smaller than) those for child participation, with confidence intervals including the null. The complete case associations were generally larger in magnitude but in the same direction as those obtained using MI; the exception to this was for asthma (Supplementary Table S6, available as Supplementary data at *IJE* online).

Impact of non-response on exposure-outcome associations

Figure 2 gives estimated associations (ORs or regression coefficients) between maternal smoking (ever/never) and outcomes derived from linked data for all individuals and for the subset who completed the questionnaire at 19 years. The associations in the subsample were all in the same direction as in the full sample and some were of a similar magnitude. The associations with offspring depression and smoking were stronger in the subsample; conversely, the associations with SEN status, asthma and attainment were attenuated in the subsample, although the confidence intervals overlapped in every case.

Discussion

We found that participation in ALSPAC is associated with a wide range of baseline sociodemographic factors; in all cases, factors suggesting greater social disadvantage were associated with lower participation. After taking account of these factors, higher educational attainment, asthma diagnosis and higher GP consultation rates and prescribed drug counts were associated with higher participation; greater school absence, special education needs status, smoking, depression and higher BMI were associated with lower participation. We also showed that some exposure-outcome estimates may be biased if restricted to current participants.

Recently, others found evidence that polygenic risk scores for education, smoking initiation, BMI and depression—as well as scores for other traits—were associated with ongoing participation in ALSPAC[3]; these associations were in the same direction as those found in the current study. Similar findings have been reported in a review of factors associated with participation in epidemiological studies[20] and in studies looking specifically at the association between psychiatric disorders and dropout in longitudinal studies.[21,22] Although some have suggested that the associations with participation would have to be quite extreme for the resulting bias to be of concern,[23] results from non-randomized studies are increasingly being combined and used to inform clinical practice.[24,25] Even a relatively small amount of bias could be problematic, particularly if similar characteristics are related to ongoing participation in different studies—results may then be reproducible from one study to another, but these results may be equally

biased. In 1998, Egger *et al*. discussed the pitfalls of combining results from observational studies. In particular, they argued that such studies may give rise to findings that are biased, and combining these (biased) estimates will result in an overall finding that is 'very precise but equally spurious'.[26] On the other hand, the factors associated with participation across studies could also differ. Knowledge of these differences could help explain heterogeneity between studies, which is useful in the context of both meta-analysis and triangulation of studies (for example, where results from one study could be compared with those from a study where the factors associated with participation differ). This suggests the need for the community to assess patterns and predictors of study participation in order to inform this type of work.

A key limitation of our study is the fact that most of the variables were incomplete. By definition, the group with complete data had higher participation rates than those with one or more missing covariates. This may mean that the observed associations with participation are themselves biased. Similarly, linked data were not available for all individuals and these individuals were not a random subsample of the cohort. Individuals who had died, withdrawn from the study after age 14, could not be traced or were flagged on the ALSPAC administrative database as being not contactable were not sent fair-processing materials and were thus not eligible to be included in the linkage. Most of those without linked education data are those who were attending an independent (fee-paying) school at the time of linkage.[5] Likewise, since the majority of primary care data come from local practices, the sample with linked GP data will mainly comprise those who have not moved out of the area. In addition, among individuals with linked primary care data, those with either low or high BMI may be more likely to have it recorded (because there may be health concerns about adolescents who appear obviously over- or underweight). This could result in bias in the estimate of the association between BMI and participation. We expect the estimates of associations with participation obtained using MI to be less biased than complete case analyses because, by including participation at each time point, all the factors from the linked data, as well as additional auxiliary variables from ALSPAC (IQ, for example) we have a better approximation to missing at random (MAR). However, the associations with participation obtained using MI may still be subject to bias.

Our findings have several implications. Statistical methods used to analyse incomplete data all make assumptions about the missing data mechanism which cannot be investigated using the incomplete data alone. For example, standard implementations of MI assume the data are MAR, conditional on the variables included in the imputation model. Similarly, a complete case analysis will generally produce unbiased estimates of the exposure-outcome association if missingness is unrelated to the outcome of interest.[27] In a previous study, we used linked educational attainment data to demonstrate that estimates of the association between duration of breastfeeding and IQ are biased due to missing data in ALSPAC, and that using attainment as an auxiliary variable in multiple regression could reduce this bias by getting closer to being missing at random. Rather than focusing on a specific outcome, in the current study we extend this work, showing that many baseline and later measures are associated with ongoing participation. This suggests that—first—a wide range of baseline covariates may need to be included in the (complete case) analysis—or imputation/weighting model—in order for the (missing data) assumptions to be met. Second, we found that many outcomes measured in adolescence were associated with participation, suggesting that a complete case analysis for such outcomes is likely to be biased. This finding is unlikely to be unique to ALSPAC, as indicated recently.[28]

Our study illustrates two key advantages of linking to external datasets. First, it is not usually possible to examine the impact of non-baseline factors on ongoing study participation because these factors are—by definition—not measured on individuals who have dropped out, whereas linkage to external datasets means later measures could be obtained on all (or a reasonably large proportion of) individuals in a study, regardless of whether they remain active participants. Second, previous work suggests that the inclusion in MI models of auxiliary variables that are reasonably highly correlated with a missing outcome variable is likely to reduce bias,[29] as it gives a closer approximation to MAR. As in the previous study, these auxiliary variables could be proxies for a missing study outcome. There are several challenges and barriers to establishing linkage to routine health and administrative data.[30] That said, it is increasingly seen to be of value in research, and many studies—both in the UK and internationally, and those across different disciplinary traditions—have or are in the process of establishing linkage to such sources.[30–32]

In conclusion, we have shown using linked data that health and educational factors are associated with ongoing participation in ALSPAC. Similar associations are likely to be present in other cohort studies; equally, specific factors may be associated with participation in one study but not others. Use of linked data will help future researchers establish whether specific analyses are likely to be biased if restricted to complete cases, and which statistical methods are likely to minimize bias. Further, knowledge of the factors associated with participation in a given study is

informative in the context of comparing and summarizing evidence across studies.

Data Availability

The authors do not have the authority to share the data that support the findings of this study, due to ALSPAC data access permissions, but any researcher can apply to use ALSPAC data, including the variables used in this investigation. Information about access to ALSPAC data is given on their website: (http://www.bristol.ac.uk/alspac/researchers/access/).

Supplementary Data

Supplementary data are available at *IJE* online.

Funding

This work was supported by the Medical Research Council (MR/L012081) and the Wellcome Trust (WT086118/Z/08/Z). The UK Medical Research Council (MRC), the Wellcome Trust and the University of Bristol currently provide core funding for ALSPAC (102215/2/13/2). Data collection is funded from a wide range of sources. This publication is the work of the authors, and R.C. will serve as guarantor for the contents of this paper. R.C. and K.T. work in the MRC Integrative Epidemiology Unit funded by the UK Medical Research Council (MC_UU_00011/3) and the University of Bristol.

Acknowledgements

We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses.

Author Contributions

R.C. and K.T. conceived of the study. R.C. conducted the statistical analyses and wrote the first draft of the manuscript. K.T. contributed to the interpretation of the results. A.B. and J.M. led the design and implementation of the ALSPAC record linkage strategy. All authors contributed to the drafting and revising of the manuscript. All authors have read and approved the final version.

Conflict of Interest

None declared.

References

1. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*, 3rd edn. Philadelphia,PA: Lippincott, Williams and Wilkins, 2008.
2. Hughes RA, Heron J, Sterne JAC, Tilling K. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *Int J Epidemiol* 2019;**48**:1294–304.
3. Taylor AE, Jones HJ, Sallis H *et al.* Exploring the association of genetic factors with participation in the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol* 2018;**47**:1207–16.
4. Martin J, Tilling K, Hubbard L *et al.* Association if genetic risk for schizophrenia with nonparticipation over time in a population-based cohort study. *Am J Epidemiol* 2016;**183**:1149–58.
5. Cornish RP, Tilling K, Boyd A, Davies A, Macleod J. Using linked educational attainment data to reduce bias due to missing outcome data in estimates of the association between the duration of breastfeeding and IQ at 15 years. *Int J Epidemiol* 2015;**44**:937–45.
6. Faris PD, Ghali WA, Brant R, Norris CM, Galbraith PD, Knudtson ML. Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses. *J Clin Epidemiol* 2002;**55**:184–91.
7. Hebert PL, Taylor LT, Wang JJ, Bergman MA. Methods for using data abstracted from medical charts to impute longitudinal missing data in a clinical trial. *Value Health* 2011;**14**:1085–91.
8. Schomaker M, Gsponer T, Estill J, Fox M, Boulle A. Non-ignorable loss to follow-up: correcting mortality estimates based on additional outcome ascertainment. *Stat Med* 2014;**33**:129–42.
9. Boyd A, Golding J, Macleod J *et al.* Cohort Profile: The 'Children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol* 2013;**42**:111–27.
10. Fraser A, Macdonald-Wallis C, Tilling K *et al.* Cohort Profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *Int J Epidemiol* 2013;**42**:97–110.
11. Cornish RP, Henderson J, Boyd AW, Granell R, Van Staa T, Macleod J. Validating childhood asthma in an epidemiological study using linked electronic patient records. *BMJ Open* 2014;**4**:e005345.
12. Cornish RP, John A, Boyd A, Tilling K, Macleod J. Defining adolescent common mental disorders using electronic primary care data: a comparison with outcomes measured using the CIS-R. *BMJ Open* 2016;**6**:e913167.
13. Atkinson MD, Kennedy JI, John A *et al.*; on behalf of the DEMISTIFY Research Group. Development of an algorithm for determining smoking status and behaviour over the life course from UK electronic primary care records. *BMC Med Inform Decis Mak* 2017;**17**:2.
14. Mukherjee M, Gupta R, Farr A *et al.* Estimating the incidence, prevalence and true cost of asthma in the UK: secondary analysis of national stand-alone and linked databases in England, Northern Ireland, Scotland and Wales—a study protocol. *BMJ Open* 2014;**4**:e006647.
15. Durrleman S, Simon R. Flexible regression models with cubic splines. *Stat Med* 1989;**8**:551–61.
16. Harrell FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression and Survival Analysis*, 2nd edn. Basel, Switzerland: Springer International Publishing, 2015.

17. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res* 2007;**16**:219–42.
18. Wechsler D. *The Wechsler Intelligence Scale for Children*, 3rd edn. San Antonio, TX: Psychological Corporation, 1991.
19. StataCorp. *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LLC, 2017.
20. Galea S, Tracy M. Participation rates in epidemiologic studies. *Ann Epidemiol* 2007;**17**:643–53.
21. Bjertness E, Sagatun Å, Green K, Lien L, Søgaard AJ, Selmer R. Response rates and selection problems, with emphasis on mental health variables and DNA sampling, in large population-based, cross-sectional and longitudinal studies of adolescents in Norway. *BMC Public Health* 2010;**10**:602.
22. de Graaf R, Bijl RV, Smit F, Ravelli A, Vollebergh WA. Psychiatric and sociodemographic predictors of attrition in a longitudinal study: The Netherlands Mental Health Survey and Incidence Study (NEMESIS). *Am J Epidemiol* 2000;**152**: 1039–47.
23. Pizzi C, De Stavola B, Merletti F *et al*. Sample selection and validity of exposure-disease association estimates in cohort studies. *J Epidemiol Community Health* 2011;**65**:407–11.
24. Sox HC, Greenfield S. Comparative effectiveness research: a report from the Institute of Medicine. *Ann Intern Med* 2009;**151**: 203–05.
25. Ijaz SI, Mischke C, Ruotsalainen J, Verbeek J, Ojajaarvi A, Neuvonen K. The use of non-randomised studies in systematic reviews of intervention effectiveness: a content analysis of Cochrane systematic reviews. *Occup Environ Med* 2013; **70**(**Suppl 1**):A60.
26. Egger M, Schneider M, Davey Smith G. Spurious precision? Meta-analysis of observational studies. *BMJ* 1998;**316**:140–44.
27. Carpenter JR, Kenward MG. *Multiple Imputation and Its Application*. Chichester, UK: Wiley, 2013.
28. Tyrrell J, Zheng J, Beaumont R. Genetic predictors of participation in optional components of UK Biobank. bioRxiv 2020. doi: 10.1101/2020.02.10.941328.
29. Cornish RP, Macleod J, Carpenter JR, Tilling K. Multiple imputation using linked proxy outcome data resulted in important bias reduction and efficiency gains: a simulation study. *Emerg Themes Epidemiol* 2017;**14**:14.
30. The Wellcome Trust. Enabling Data Linkage to Maximise the Value of Public Health Research Data: Full Report. London: Wellcome Trust, 2015.
31. Medical Research Council. Maximising the Value of UK Population Cohorts: MRC Strategic Review of the Largest UK Population Cohort Studies. London: MRC, 2014.
32. Davis-Kean P, Chambers RL, Davidson LL, Kleinert C, Ren Q, Tang S. *Longitudinal Studies Strategic Review: 2017 Report to the Economic and Social Research Council*. London: ESRC, 2018.